第46卷 第10期
2023年10月

干旱区地理
ARID LAND GEOGRAPHY

Vol. 46 No. 10
Oct. 2023

基于XGBoost模型的湟水流域耕地土壤养分遥感反演

刘尊方[1]， 雷浩川[1]， 盛海彦[2]

（1. 青海大学地质工程系，青海 西宁 810016；2. 青海大学农牧学院，青海 西宁 810016）

摘 要：湟水流域是河湟谷地重要的组成部分，协同环境因素预测土壤养分空间分布对农业土壤养分管理尤为重要。土壤养分反演研究中对于参数对模型结果的影响和模型适用性的研究较少。选取研究区地形因子、土壤pH及光谱反射率共28个因子，结合贝叶斯优化算法构建人工神经网络（ANN）、支持向量机（SVM）和极端梯度提升（XGBoost）3种机器学习模型预测耕地土壤养分空间分布，计算决定系数（R^2）、均方根误差（RMSE）和相对分析误差（RPD）评价3种模型的精度。结果表明：（1）基于贝叶斯优化超参数的XGBoost模型对全氮（TN）含量预测精度优于其他模型（R^2=0.893，RMSE=0.359，RPD=2.470），预测土壤有机质（SOM）、速效磷（AP）和速效钾（AK）含量时，XGBoost模型验证集R^2分别为0.801、0.509、0.442。（2）对比3种模型的寻优次数和误差发现，BOA-XGBoost模型参数优化次数少、效率高，具有更好的鲁棒性。对于不同的养分，ANN和SVM模型预测精度存在差异，SVM模型预测SOM含量时精度更高（RPD=1.580），而ANN模型预测TN时精度最佳（RPD=2.460）。基于贝叶斯算法进行超参数优化构建的XGBoost模型预测精度高，可以达到良好的预测效果，可为湟水流域精准农业施肥提供参考。

关 键 词：土壤养分；XGBoost；空间分布；环境因子；湟水流域

文章编号：1000－6060(2023)10－1643－11(1643～1653)

土壤有机质（SOM）、全氮（TN）、速效磷（AP）和速效钾（AK）等养分是植被生长所必需的[1]，同时也是评价土壤肥力水平的重要指标。肥沃的土壤可以给作物提供丰富的营养物质促进作物生长，而贫瘠的土壤限制作物生长，进而影响到产量。因此，对耕地土壤养分进行空间分布预测和制图，是实现可持续农业的基本要求[2]，同时也对指导变量施肥和耕地土壤质量提升具有重要意义。

基于地统计学分析和GIS技术的养分含量空间分布模拟精度取决于土壤样品采样密度，预测过程没有考虑土壤养分空间分布与环境因素和土壤自身属性的密切联系[3-4]；传统土壤养分含量检测耗时长且所需费用高，不适合大范围养分含量空间分布制图。随着遥感技术的发展，在预测养分含量空间分布和评估土壤肥力方面得到了广泛应用，实现了大范围、高效率的土壤养分含量空间分布监测[5]。该技术在地面少量实测数据的基础上，通过分析光谱信息并提取敏感光谱波段构建反演模型，实现大面积养分含量空间分布预测。目前，已有多种机器学习模型结合遥感影像进行土壤养分空间分布的反演。李冠稳等[6]以湟水流域不同粒径SOM为研究对象，构建的SVM模型预测精度比偏最小二乘模型精度高，验证集相对分析误差大于2，适用于湟水流域SOM含量预测；雷浩川等[7]以Landsat遥感影像为数据源，将经过倒数处理的波段反射率作为模型自变量，构建神经网络模型反演湟水流域大通县TN含量空间分布格局，模型决定系数（R^2）为0.792，均方根误差（RMSE）为0.246；此外，肖云飞[8]和李冠稳等[9]基于随机森林模型反演湟水流域土壤SOM含量，模型相对分析误差（RPD）分别达到了4.229和

收稿日期：2023-01-19； **修订日期**：2023-04-17
基金项目：国家自然科学基金项目（U20A20115）；青海大学创新创业工坊项目（GF-20230005）资助
作者简介：刘尊方（1999-），男，硕士研究生，主要从事土壤定量遥感等方面的研究. E-mail: lzunfang@163.com
通讯作者：雷浩川（1973-），男，博士，讲师，主要从事遥感与GIS应用等方面的研究. E-mail: 56242188@qq.com

4.700。除上述模型外,极限学习机[10]、多元回归[11-12]、地理加权回归克里格[13]等模型的预测精度较高,在土壤养分预测研究中得到了广泛应用。

根据目前的研究现状,线性和非线性模型均用于土壤养分含量空间分布的研究中,然而机器学习模型精度明显比线性模型高,因此适合于土壤养分反演研究,可以提高土壤养分空间分布预测的精度[14]。土壤养分空间分布受到多种因素的影响[15],此外,机器学习模型参数设置会影响养分含量预测精度,而在已有研究中模型参数的选择多是经过试验手动调整而确定,因此在参数选择上具有偶然性。贝叶斯优化算法(BOA)是一种高效的全局智能优化算法,较其他优化算法寻求最优解的速度更快,但基于该算法选取模型参数多见于土壤盐分含量反演研究中,如杨练兵等[16]基于BOA优化随机森林模型参数反演土壤盐分;Wang等[17]使用BOA优化Light-GBM模型,研究结果精度高。然而在土壤养分空间分布研究中,运用BOA优化参数的研究尚不多见。为此,本文以青海省湟水流域为研究区,基于相关性分析选取建模因子,使用Landsat 8多光谱遥感影像和野外实测耕地土壤养分含量数据,构建人工神经网络(ANN)、支持向量机(SVM)和极端梯度提升(XGBoost)3种模型用于耕地土壤养分空间反演;利用BOA选择模型最优超参数,提高模型预测精度,构建适合湟水流域耕地养分空间预测的最优模型,为指导变量施肥和精准农业有效实施提供有利基础和技术支撑。

1 研究区概况

湟水河为黄河上游的一级支流,是黄河重要的水源地和生态屏障,同时也是青藏高原和黄土高原的过渡带[18-19]。湟水流域(36°02′~37°28′N,100°42′~103°04′E)位于青海省东北部,发源于青海省海晏县,流经西宁市和海东市,在甘肃省永登县汇入黄河(图1)。湟水流域作为河湟谷地重要的组成部分,在青海省工农业发展中占有主导位置。湟水流域属于高原大陆性气候,海拔较高,地势呈西高东低趋势,地形变化较为复杂;流域内大部分区域为干旱半干旱地区,全年干旱少雨,年平均气温低于8 ℃。流域内主要种植春小麦、油菜和青稞等农作物。

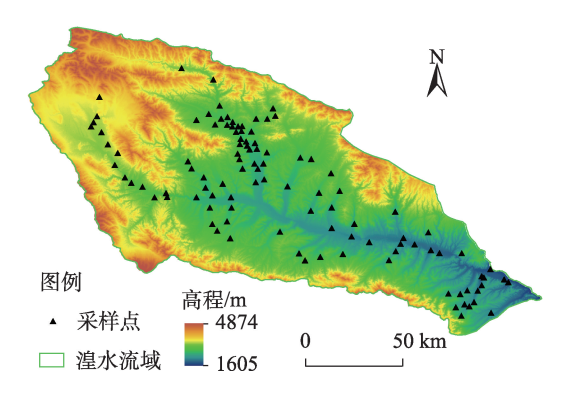


图1 研究区土壤采样点分布

Fig. 1 Distribution of soil sampling points in the study area

2 数据与方法

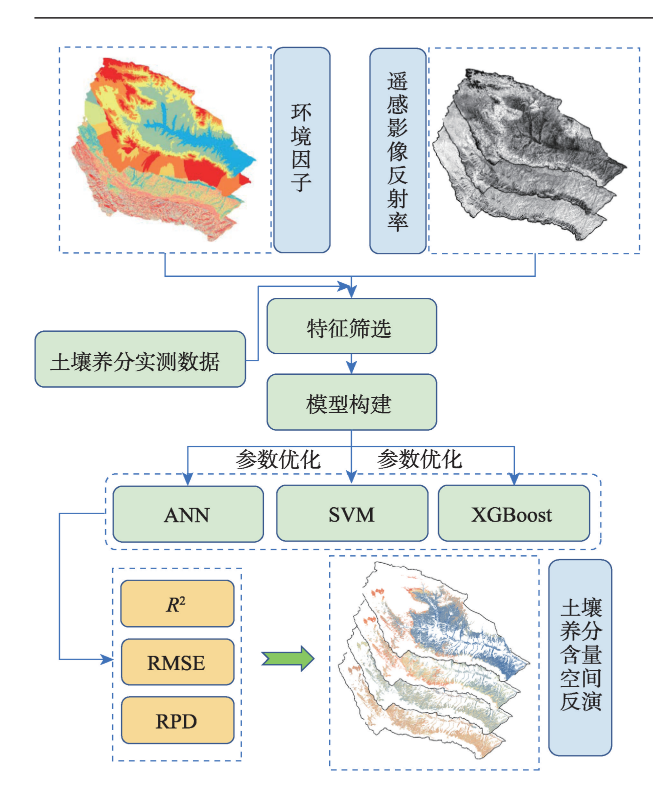
2.1 土壤样本采集及数据来源

本文综合考虑研究区地形地势、采样难易程度和交通便利情况,在2021年春耕前选取无农作物残留地或裸地采集110个耕地土壤样品,同时利用手持GPS接收机记录采样点经纬度信息,并详细记录采样点附近其他环境信息;在GPS记录点10 m直径范围内,采集5个耕地土壤样品,采样深度为0~20 cm,最终测定1个样点养分含量的土壤样品由5个位置采集样品混合而成。采集的土壤样品经自然风干去除杂质、研磨过筛后在实验室测定SOM、TN、AP和AK 4种土壤养分含量及土壤pH。数据预处理过程中,由于异常值而剔除3个土壤样品,实际用于研究的耕地土壤样品数为107个。

DEM数据来源于地理空间数据云(https://www.gscloud.cn/),分辨率为30 m;Landsat 8 OLI数据来源于USGS(https://earthexplorer.usgs.gov/),成像时间为2021年4月,云覆盖量均小于5%;并对获取的影像进行大气校正、镶嵌和裁剪预处理操作。利用Arc-GIS提取研究区坡度、坡向、地形起伏度、平面曲率、剖面曲率和地形湿度指数6种地形因子。

2.2 研究方法

本文技术路线如图2所示。首先基于Landsat 8影像和DEM数据提取土壤养分建模因子,结合Pearson相关分析按相关系数绝对值大小筛选因子。然后构建ANN、SVM、XGBoost 3种耕地土壤养分反演模型,并结合实测数据和建模因子对模型参数进行贝叶斯优化,对比3种模型反演精度,选取最优模型反演湟水流域土壤养分含量空间分布。



注：ANN为人工神经网络；SVM为支持向量机；XGBoost为极端梯度提升；R^2为决定系数；RMSE为均方根误差；RPD为相对分析误差。下同。

图2　养分反演技术路线图

Fig. 2　Flowchart of the nutrient inversion and analysis

2.2.1　*反演模型*　ANN是模拟人脑神经网络信息传递而提出的算法，该模型通过网络中神经元的相互作用来处理模糊、非线性的信息，在土壤养分反演研究中应用广泛。模型由输入层、隐藏层和输出层构成，各层之间由大量神经元通过权重而连接；隐藏层可以有多层，输入层和输出层各包含1层[7]，输入层即为环境因子和波段反射率，输出层为4种养分含量值。模型最大迭代次数、隐藏层神经元个数以及激活函数由BOA确定。

SVM是基于统计学理论提出的机器学习算法，算法可以很好地解决小样本、非线性问题[18]，其核心技术在于核函数的选择，在解决实际问题时可根据研究问题的不同合理选择核函数。该算法中共有4种核函数，分别是多项式核函数、Sigmoid核函数、线性核函数和径向基函数(RBF)，在土壤测绘研究中RBF是最常用的核函数[20]。本文以RBF为基函数构建SVM模型，由BOA确定模型的惩罚系数和核参数。

XGBoost是一种新兴集成学习算法，算法核心思想在于每次迭代生成一棵树，即学习一个新函数拟合上次迭代预测值残差，以生成的树为基础，再训练出新的树。该算法在梯度提升树基础上，保留了更多目标函数信息，提高了模型训练速度，并在目标函数中加入正则项，控制模型的复杂程度，降低过拟合现象的发生[21-22]，有效改进了梯度提升树算法。模型学习率、树的深度等参数经BOA确定。

BOA与传统随机搜索和网格搜索算法相比，该算法会综合考虑已经搜索过参数对模型的表现，减少优化迭代次数，提高优化效率。算法流程如下：

（1）随机选择若干组参数x，添加数据训练模型获取对应指标y；

（2）通过代理函数拟合x与y；

（3）使用采集函数选择最佳参数组合x^*；

（4）将x^*代入模型获取新的y，重复第(2)步，直到最优参数被确定。

2.2.2　*评价指标*　研究采用R^2、RMSE和RPD评价ANN、SVM和XGBoost 3种耕地土壤养分反演模型的精度。RMSE越小、R^2越大，表示模型的预测值与养分实测值偏差小、相关性高，说明模型预测精度高，可以更好地反映研究区耕地土壤养分含量的空间分布特征；当模型RPD>2时预测精度高，1.4<RPD<2时预测精度较好，RPD<1.4时预测精度低。精度评价指标计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2} \qquad (1)$$

$$R^2=\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \qquad (2)$$

$$\mathrm{RPD}=\frac{\mathrm{SD}}{\mathrm{RMSE}} \qquad (3)$$

式中：n为采集土壤样本总数；y_i为养分含量实测值；$\hat{y}_i$为模型预测养分含量值；$\bar{y}$为养分平均值；SD为验证集标准差。

3　结果与分析

3.1　土壤养分含量常规统计分析

由湟水流域土壤养分描述性统计分析结果可知(表1)，湟水流域SOM含量介于4.310~59.789 g·kg^{-1}，TN含量介于0.372~5.874 g·kg^{-1}之间，AP含量介于0.001~0.162 g·kg^{-1}之间，AK含量介于0.048~0.488 g·kg^{-1}之间；SOM、TN、AP和AK平均含量分别为28.377 g·kg^{-1}、

表1 湟水流域土壤养分常规统计分析结果

Tab. 1 Routine statistical analysis results of soil nutrients in the Huangshui River Basin

土壤养分	样点数/个	最小值/g·kg^{-1}	最大值/g·kg^{-1}	平均值/g·kg^{-1}	标准差/g·kg^{-1}	变异系数/%
SOM	107	4.310	59.789	28.377	11.168	39.356
TN	107	0.372	5.874	1.232	0.856	69.481
AP	107	0.001	0.162	0.055	0.032	58.182
AK	107	0.048	0.488	0.204	0.087	42.647

注：SOM、TN、AP和AK分别为土壤有机质、全氮、速效磷和速效钾。下同。

1.232 g·kg^{-1}、0.055 g·kg^{-1}和0.204 g·kg^{-1}。变异系数用来反映土壤养分含量的离散程度，经统计分析后，4种土壤养分的变异系数分别为39.356%、69.481%、58.182%和42.647%，4种养分均处于中等变异程度。

3.2 土壤养分与地形因子、波段反射率相关性分析

以Landsat 8影像波段原始反射率(b1~b7)、经过数学变换的反射率(对数和倒数处理)、6种地形因子及土壤pH，共计28个因子作为模型的输入变量。由环境变量与4种耕地土壤养分含量Pearson相关性分析结果(表2)可见，研究区SOM含量与高程和土壤pH值呈极显著相关关系($P<0.01$)，说明区域的海拔越高、pH值越小，SOM的含量越高；SOM含量与坡向、剖面曲率、地形湿度指数、地形起伏度之间呈显著相关性($P<0.05$)，而与坡度和平面曲率相关性不高；TN与SOM的相关性较高，都与环境因子呈现相似的相关性，研究区海拔较高地区SOM和TN含量较高；AP与研究区坡向、平面曲率和地形起伏度之间呈现显著相关关系，与其他因子的相关性不显著；当地形的起伏变化越大，研究区的AP含量就越低；AK与坡度、pH呈现显著负相关性，坡度越大、土壤酸性越弱，研究区AK含量越低，这与Komolafe等[23]和代子俊[24]的研究结果一致。

由耕地土壤养分含量和反射率Pearson相关分析结果(表3)可见，SOM和原始波段都呈现相关关系，而且相比原始波段反射率，经过数学变换处理

表2 土壤养分与环境变量间的相关性分析

Tab. 2 Correlation analysis between soil nutrients and environmental variables

土壤养分	高程	坡向	坡度	平面曲率	剖面曲率	地形湿度指数	地形起伏度	pH
SOM	0.422**	0.223*	−0.022	0.101	0.052*	0.010*	0.095*	−0.338**
TN	0.595**	0.238*	−0.174*	−0.015	0.089*	0.052*	0.276**	−0.485**
AP	−0.048	−0.066*	−0.028	0.080*	−0.036	−0.048	−0.098*	0.026
AK	0.052	0.041	−0.080*	0.029	0.015	−0.015	−0.013	−0.106*

注：**、*分别表示在$P<0.01$、$P<0.05$水平上显著。下同。

表3 土壤养分与波段反射率之间的相关性分析

Tab. 3 Correlation analysis between soil nutrients and band reflectance

波段	SOM	TN	AP	AK	波段	SOM	TN	AP	AK
b1	−0.390**	−0.343**	0.118*	−0.039	lg5	−0.283**	−0.158	0.088	−0.034
b2	−0.391**	−0.334**	0.108*	−0.040	lg6	−0.259**	−0.185	0.082	−0.125*
b3	−0.393**	−0.311**	0.083	−0.070	lg7	−0.251**	−0.247*	0.069	−0.169*
b4	−0.379**	−0.280**	0.085	−0.084*	1/b1	0.278**	0.424**	−0.134	0.046
b5	−0.267**	−0.084	0.068	−0.041	1/b2	0.334**	0.485**	−0.151*	0.034
b6	−0.194*	−0.018	0.033	−0.120	1/b3	0.376**	0.491**	−0.142*	0.044
b7	−0.174*	−0.073	0.008	−0.180*	1/b4	0.385**	0.489**	−0.143*	0.063
lg1	−0.387**	−0.495**	0.150*	−0.024	1/b5	0.300**	0.247*	−0.112	0.029
lg2	−0.392**	−0.452**	0.139*	−0.019	1/b6	0.317**	0.361**	−0.128	0.119*
lg3	−0.394**	−0.407**	0.113	−0.051	1/b7	0.316**	0.422**	−0.126	0.139*
lg4	−0.387**	−0.386**	0.114	−0.071					

注：b1~b7为波段原始反射率；lg1~lg7为经对数处理的反射率；1/b1~1/b7为经倒数处理的反射率。下同。

之后的反射率与SOM的相关性提高，如b5原始反射率与SOM的相关系数（r）=−0.267，经过倒数处理之后，r=0.300；TN与波段反射率相关性与SOM一致，TN和lg1的相关性最高（r=−0.495）；总体而言AP和AK与反射率的相关关系极低，只有个别波段通过显著性检验$P<0.05$，如1/b2和AP之间呈现显著相关关系（r=−0.151，$P<0.05$），AK与lg7相关关系最佳（r=−0.169，$P<0.05$）。

3.3 反演模型构建及优化

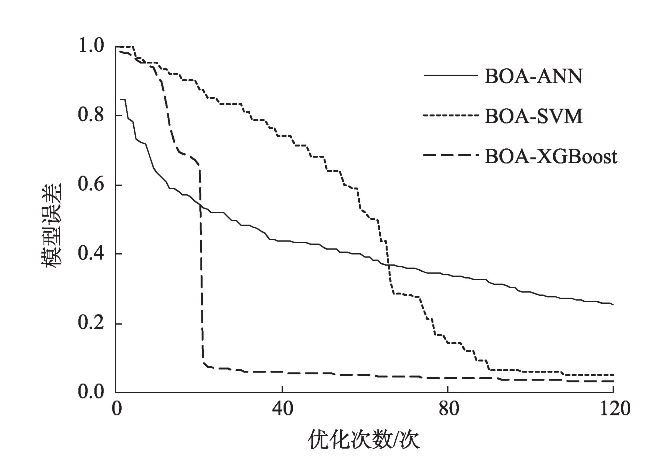
土壤养分空间分布受到人为因素（施肥、灌溉等）和自然因素（地形、土壤类型等）的共同影响[25-26]，建模因子的选取影响反演模型预测精度，因此在兼顾研究区地形、影像波段原始反射率及数学变换反射率3类因子的前提下，依据相关性分析结果以及研究区特点，优先选取与4种耕地土壤养分呈现极显著相关关系的环境因子、原始反射率和经过数学变换之后的反射率作为建模因子。为避免选取的建模因子集中在同一类，通过分段极大值法选取各养分建模因子，即从上述的3类因子中选相关性高的因子作为该养分反演模型的建模因子，由于养分含量及养分自身性质与3类因子的相关性不同，建模因子的选取存在差异。各土壤养分建模因子选取结果如表4所示。

表4 土壤养分建模因子

Tab. 4 Soil nutrient modeling factors

养分	建模因子
SOM	高程、土壤pH、地形起伏度、lg3、1/b4
TN	高程、坡向、剖面曲率、地形起伏度、土壤pH、lg1、1/b3
AP	坡向、平面曲率、地形起伏度、lg1、1/b2
AK	坡度、地形湿度指数、土壤pH、b7

利用上述建模因子和实测耕地养分含量数据，构建反演模型，并基于BOA对模型参数进行优化，选取最优的参数组合使得模型预测误差最小。随着优化次数的增加，模型误差趋于稳定，此时模型各参数值即为最优参数组合。通过比较3种模型的优化迭代次数以及模型预测误差（图3），发现BOA-XGBoost模型寻找最优参数组合所需的优化次数最少，BOA-ANN模型最终预测土壤养分空间分布的误差较大。



注：BOA-ANN、BOA-SVM和BOA-XGBoost分别为经贝叶斯优化的人工神经网络、支持向量机和极端梯度提升。

图3 模型优化结果

Fig. 3 Model optimization results

贝叶斯优化模型参数过程中，采用5折交叉验证法确定机器学习模型最终的参数值，经贝叶斯优化3种模型的最优参数组合如表5所示。

3.4 模型精度比较

基于模型最优超参数值，随机选取70%的样本分别构建反演模型，并以剩余样本对模型进行验

表5 3种模型最优超参数组合

Tab. 5 Optimal hyperparametric combination of the three models

模型	参数	土壤养分			
		SOM	TN	AP	AK
ANN	隐藏层节点/个	80	100	120	120
	最大迭代次数/次	1000	1500	500	500
	激活函数	S型函数	S型函数	修正线性单元函数	修正线性单元函数
SVM	惩罚系数	3.12	2.50	0.35	1.10
	核参数	0.42	0.73	0.91	0.92
XGBoost	树的最大深度	4	5	4	4
	最大树数目	50	30	75	96
	学习率	0.03	0.07	0.05	0.04
	最小叶节点样本权重和	4	4	2	1

注：ANN、SVM、XGBoost分别为人工神经网络、支持向量机、极端梯度提升。下同。

证，选取最优反演模型用于湟水流域耕地土壤养分空间预测。表6为各模型预测精度比较，可以看出，XGBoost模型反演精度明显高于ANN和SVM模型。SOM和TN验证集R^2分别为0.801、0.893，RPD均大于2，具有较高的预测精度，与ANN模型预测SOM和TN时相比，R^2分别提高了0.158、0.090；而与SVM模型相比，R^2分别提高了0.169、0.158；对于AP和AK的预测，3种模型验证集R^2均不高，但相比于其他模型，XGBoost模型的R^2仍然最大，RMSE最小，精度显著优于ANN和SVM模型。

3.5 土壤养分空间预测

基于反演模型对4种耕地土壤养分进行空间预测，并绘制空间预测分布图(图4)。从预测图中可以看出，不同模型对于相同养分含量的反演存在差异；SOM和TN之间相关性高，空间分布总体变化趋势相似，即西北分布含量高，东南部含量较低；利用XGBoost模型反演研究区SOM含量在7.91~55.74 g·kg^{-1}之间，而TN含量反演最低值为0.42 g·kg^{-1}，最高值为5.14 g·kg^{-1}。AP和AK总体含量较低，其空间分布变化不明显，而且AP和AK含量之间呈现负相关性，AP含量越高，则AK含量越低。XGBoost模型预测AP含量范围为0.04~0.11 g·kg^{-1}，而AK含量范围为0.07~0.42 g·kg^{-1}。综上所述，XGBoost模型反演值更接近实测数据范围，基本可以反演出研究区实际耕地土壤养分含量的空间分布状态。

土壤养分的分布具有空间自相关性和异质性[24-26]。以SOM为例，选取子区域绘制SOM局部空间分布特征(图5)，来比较3种模型的适用性。发现3种模型对SOM空间分布特征反演存在较大的差异；对比真彩色合成遥感影像，XGBoost模型预测SOM分布细节更清晰，能更好反映SOM空间分布。

4 讨论

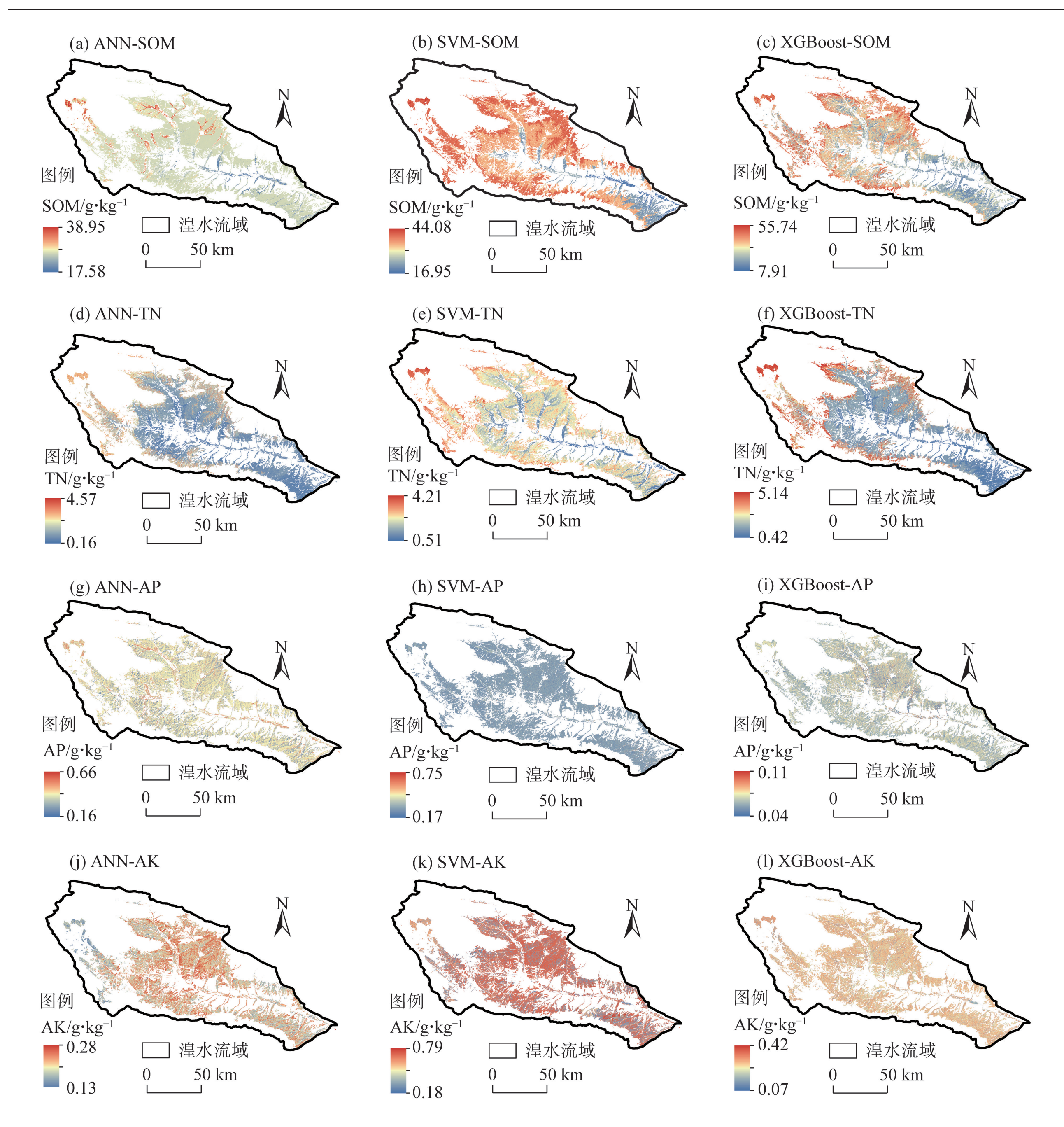
XGBoost模型是一种具有训练速度快、拟合精度高的新兴集成学习算法，在盐分含量反演[27]、冬小麦全氮含量反演[21]等领域得到了很好的应用，而在土壤养分反演研究中应用较少。本研究利用BOA-XGBoost模型反演耕地土壤养分含量，分析了XGBoost模型在湟水流域耕地土壤养分反演中的适用性。研究发现不同算法对同一养分的建模效果不同，其中XGBoost模型的预测精度最高；SVM和ANN 2种模型对于不同的养分，其预测精度存在差异，前者预测SOM时精度高(RPD=1.580)，后者则对TN预测效果好(RPD=2.460)。其他学者也采用不同的方法对湟水流域养分进行估测，如胡亚男等[28]构建PLSR模型预测湟水流域SOM，模型验证集R^2=0.690、RPD=1.820，其精度低于本文中XGBoost模型的精度(R^2=0.801，RMSE=4.321，RPD=2.152)；张欢等[13]基于地统计学方法模拟SOM含量、刘尊方等[26]以空间插值方法绘制的湟水流域SOM含量空间分布趋势与本文构建的最优模型反演得到的空间分布一致。空间插值方法得到的养分空间分布在细节上得不到体现，遥感技术结合机器学习方法可以

表6 不同养分反演模型精度比较

Tab. 6 Precision comparison of different nutrient inversion models

土壤养分	模型	建模集		验证集		
		R^2	RMSE	R^2	RMSE	RPD
SOM	ANN	0.753	5.363	0.643	7.401	1.214
	SVM	0.822	5.122	0.632	5.648	1.580
	XGBoost	0.910	3.791	0.801	4.321	2.152
TN	ANN	0.885	0.245	0.803	0.306	2.460
	SVM	0.871	0.415	0.735	0.252	1.886
	XGBoost	0.958	0.235	0.893	0.359	2.470
AP	ANN	0.491	0.040	0.382	0.030	1.002
	SVM	0.468	0.023	0.441	0.029	1.213
	XGBoost	0.692	0.022	0.509	0.026	1.210
AK	ANN	0.514	0.063	0.419	0.064	1.321
	SVM	0.486	0.061	0.354	0.072	1.260
	XGBoost	0.692	0.043	0.442	0.055	1.274

注：R^2、RMSE、RPD分别为决定系数、均方根误差、相对分析误差。下同。



注：SOM、TN、AP和AK分别为土壤有机质、全氮、速效磷和速效钾。

图4 3种模型反演的土壤养分空间分布

Fig. 4 Spatial distributions of soil nutrients by three models

弥补这一缺点，但其反演精度取决于影像的空间分辨率[29-30]。

已有研究表明，仅通过环境变量预测土壤养分含量时精度较低[31]，而基于遥感影像波段反射率构建模型时，又忽略了环境因素对养分含量空间分布的影响，反演结果不可靠。张欢等[13]的研究发现湟水流域SOM空间分布受到地形、土壤类型、土地利用类型和施肥等多方面的影响。本文也分析了养分和环境因子之间的相关性，发现高程、坡向、地形起伏度是影响养分空间分布的主要因素。Liu等[14]仅通过多光谱波段反射率构建TN反演模型，忽略了环境因子的影响，模型精度较低(R^2=0.676)；李莹莹等[32]的研究在反射率基础上加入环境因子之后，模型预测精度提高。在选取特征波段时，可对原始光谱反射率进行数学变换，这样可以有效提高反射率与土壤养分之间的相关性，有利于建立精度更高

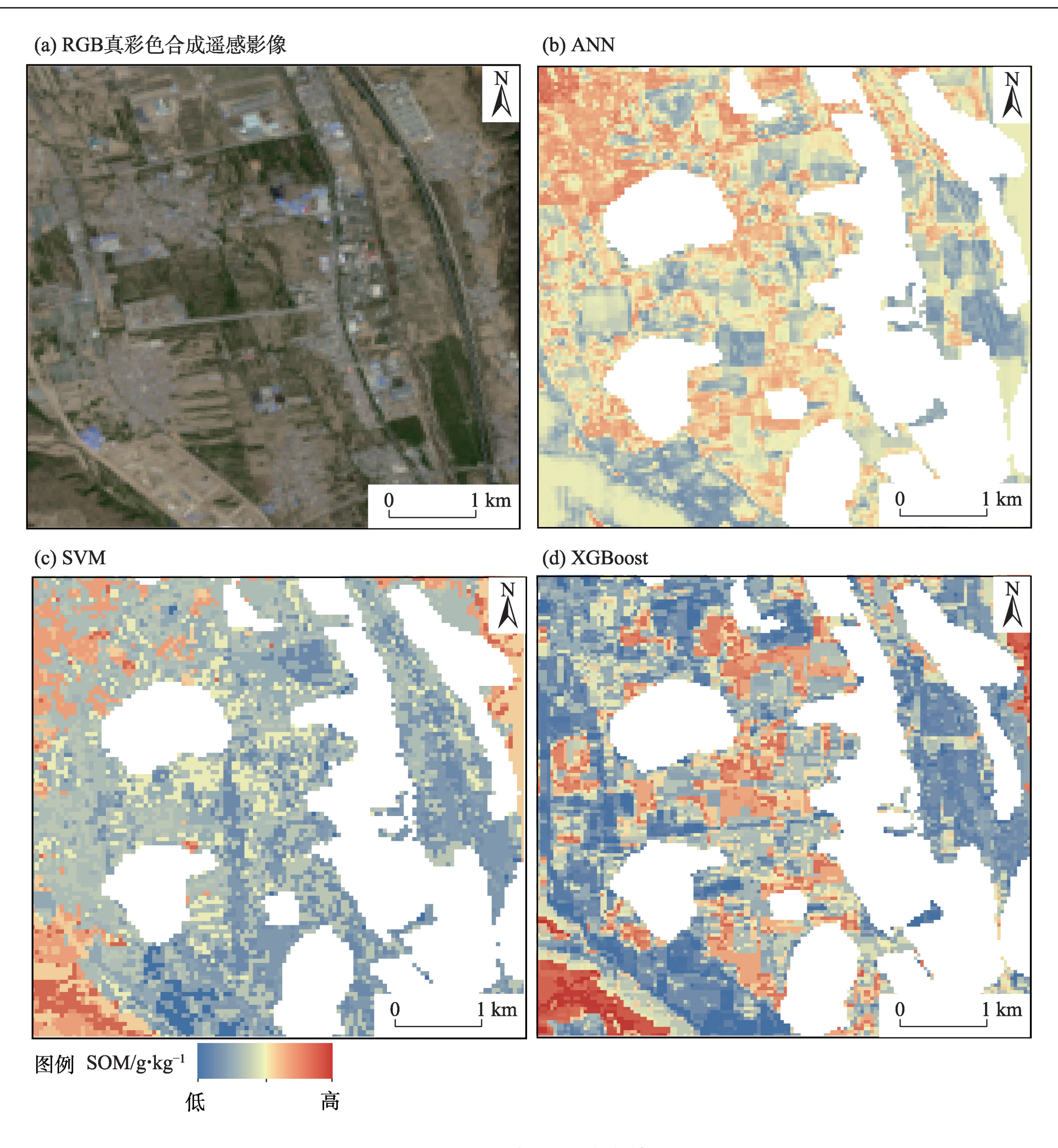


图5 SOM局部空间分布特征

Fig. 5 Local spatial distribution characteristics of SOM

的反演模型，本文对Landsat 8原始影像b3反射率经倒数处理之后，相关系数由原来的-0.311提高到0.491，相关系数绝对值提高了57.88%；孙铭岳[33]基于多源遥感数据反演SOM和盐分的研究中，也表明经过数学变换之后的反射率与有机质的相关性高，杜军等[34]的研究也得到类似的结果。

众多因素共同决定着土壤养分的空间分布，而本文选取的辅助因子有限，今后工作将侧重于环境和人为因子协同条件下土壤养分的空间反演，以及其他优化算法对模型参数选择的优缺点；为准确预测微量养分含量空间分布，考虑选择高光谱遥感影像数据以提取与养分显著相关的波段，提高预测能力，为湟水流域精准农业施肥提供参考。

5 结 论

（1）湟水流域土壤有机质（SOM）、全氮（TN）、速效磷（AP）和速效钾（AK）4种养分均处于中等变异程度，其中TN的变异程度最大（69.481%）；4种养分的含量均在中等水平。

（2）结合湟水流域6种环境因子、波段反射率和衍生反射率构建反演模型，比较ANN、SVM和XGBoost 3种机器学习模型的预测精度，发现ANN和SVM模型的预测能力较低，XGBoost模型预测TN含量精度最高（R^2=0.893，RMSE=0.359，RPD=2.470），反演SOM含量空间分布精度次之，验证集R^2=0.801，

RMSE=4.321，RPD=2.152。最终绘制的SOM和TN空间分布趋势保持一致，总体呈现西北高东南低，且中部含量较低；AP和AK含量较低，变化趋势不明显。

（3）相比ANN和SVM模型，贝叶斯优化XGBoost模型反演湟水流域土壤养分空间分布格局精度高，可以很好地体现土壤养分局部分布特征，更适合应用于土壤养分反演研究中。

参考文献（References）

[1] Wang R M, Zou R Y, Liu J M, et al. Spatial distribution of soil nutrients in farmland in a hilly region of the Pearl River Delta in China based on geostatistics and the inverse distance weighting method[J]. Agriculture, 2021, 11(1): 50, doi: 10.3390/agriculture11010050.

[2] Alemu L, Mesfin B. Performance of mid infrared spectroscopy to predict nutrients for agricultural soils in selected areas of Ethiopia [J]. Heliyon, 2022, 8(3): e09050, doi: 10.1016/j.heliyon.2022.e09050.

[3] 刘焕军, 张美薇, 杨昊轩, 等. 多光谱遥感结合随机森林算法反演耕作土壤有机质含量[J]. 农业工程学报, 2020, 36(10): 134-140. [Liu Huanjun, Zhang Meiwei, Yang Haoxuan, et al. Invertion of cultivated soil organic matter content combining multi-spectral remote sensing and random forest algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(10): 134-140.]

[4] 郑淼, 王翔, 李思佳, 等. 黑土区土壤有机质和全氮含量遥感反演研究[J]. 地理科学, 2022, 42(8): 1336-1347. [Zheng Miao, Wang Xiang, Li Sijia, et al. Remote sensing inversion of soil organic matter and total nitrogen in black soil region[J]. Scientia Geographica Sinica, 2022, 42(8): 1336-1347.]

[5] Gulhane V A, Rode S V, Pande C B. Wavelet for predicting soil nutrients using remotely sensed satellite images[J]. International Journal of Computer Applications, 2017, 174(4): 35-38.

[6] 李冠稳, 高小红, 杨灵玉, 等. 不同粒径土壤有机质含量可见光-近红外光谱估算研究——以湟水流域为例[J]. 土壤通报, 2017, 48(6): 1360-1370. [Li Guanwen, Gao Xiaohong, Yang Lingyu, et al. Estimating soil organic matter contents from different soil particle size using visible and near-infrared reflectance spectrum: A case study of the Huangshui Basin[J]. Chinese Journal of Soil Science, 2017, 48(6): 1360-1370.]

[7] 雷浩川, 刘尊方, 于晓晶, 等. 基于Landsat 5影像的青海省大通县土壤表层全氮空间格局反演[J]. 青海大学学报, 2021, 39(6): 79-86. [Lei Haochuan, Liu Zunfang, Yu Xiaojing, et al. Spatial pattern inversion of soil surface total nitrogen in Datong County of Qinghai Province based on Landsat 5 image[J]. Journal of Qinghai University, 2021, 39(6): 79-86.]

[8] 肖云飞. 青海省土壤有机质、全碳、全氮高光谱遥感估算研究[D]. 西宁: 青海师范大学, 2019. [Xiao Yunfei. Estimation of soil organic matter, total carbon and total nitrogen by hyperspectral remote sensing in Qinghai Province[D]. Xining: Qinghai Normal University, 2019.]

[9] 李冠稳, 高小红, 肖能文, 等. 基于sCARS-RF算法的高光谱估算土壤有机质含量[J]. 发光学报, 2019, 40(8): 1030-1039. [Li Guanwen, Gao Xiaohong, Xiao Nengwen, et al. Estimation soil organic matter contents with hyperspectra based on sCARS and RF algorithms[J]. Chinese Journal of Luminescence, 2019, 40(8): 1030-1039.]

[10] 杨晓宇, 包妮沙, 曹粤, 等. 基于无人机成像光谱技术的农田土壤养分估测及制图[J]. 地理与地理信息科学, 2021, 37(5): 38-45. [Yang Xiaoyu, Bao Nisha, Cao Yue, et al. Estimation and mapping of soil nutrient in farmland based on UAV imaging spectrometry[J]. Geography and Geo-information Science, 2021, 37(5): 38-45.]

[11] 郑曼迪, 熊黑钢, 乔娟峰, 等. 基于高光谱的不同人类干扰程度下荒漠土壤有机质含量估算模型[J]. 干旱区地理, 2018, 41(2): 384-392. [Zheng Mandi, Xiong Heigang, Qiao Juanfeng, et al. Hyperspectral based estimation model about organic matter in desert soil at different levels of human disturbance[J]. Arid Land Geography, 2018, 41(2): 384-392.]

[12] Miran N, Rasouli Sadaghiani M H, Feiziasl V, et al. Predicting soil nutrient contents using Landsat OLI satellite images in rain-fed agricultural lands, northwest of Iran[J]. Environmental Monitoring and Assessment, 2021, 193(9): 607, doi: 10.1007/s10661-021-09397-0.

[13] 张欢, 高小红. 复杂地形区土壤有机质空间变异性分析及制图[J]. 水土保持研究, 2020, 27(5): 93-100. [Zhang Huan, Gao Xiaohong. Analysis of spatial variability and mapping of soil organic matter contents in complex terrain areas[J]. Research of Soil and Water Conservation, 2020, 27(5): 93-100.]

[14] Liu Z F, Lei H C, Lei L, et al. Spatial prediction of total nitrogen in soil surface layer based on machine learning[J]. Sustainability, 2022, 14(19): 11998, doi: 10.3390/su141911998.

[15] Dharumarajan S, Lalitha M, Niranjana K, et al. Evaluation of digital soil mapping approach for predicting soil fertility parameters: A case study from Karnataka Plateau, India[J]. Arabian Journal of Geosciences, 2022, 15(5): 386, doi: 10.1007/s12517-022-09629-8.

[16] 杨练兵, 陈春波, 郑宏伟, 等. 基于优化随机森林回归模型的土壤盐渍化反演[J]. 地球信息科学学报, 2021, 23(9): 1662-1674. [Yang Lianbing, Chen Chunbo, Zheng Hongwei, et al. Retrieval of soil salinity content based on random forests regression optimized by Bayesian optimization algorithm and genetic algorithm[J]. Journal of Geo-information Science, 2021, 23(9): 1662-1674.]

[17] Wang L Y, Hu P, Zheng H W, et al. Integrative modeling of heterogeneous soil salinity using sparse ground samples and remote sensing images[J]. Geoderma, 2023, 430: 116321, doi: 10.1016/j.geoderma.2022.116321.

[18] Liu F, Qin T L, Yan D H, et al. Classification of instream ecologi-

cal water demand and crucial values in a semi-arid river basin[J]. Science of the Total Environment, 2020, 712: 136409, doi: 10.1016/j.scitotenv.2019.136409.

[19] Dong B Q, Qin T L, Wang Y, et al. Spatiotemporal variation of nitrogen and phosphorus and its main influencing factors in Huangshui River Basin[J]. Environmental Monitoring and Assessment, 2021, 193(5): 292, doi: 10.1007/s10661-021-09067-1.

[20] Ruhollah T M, Ram N, Sood K, et al. Artificial bee colony feature selection algorithm combined with machine learning algorithms to predict vertical and lateral distribution of soil organic matter in South Dakota, USA[J]. Carbon Management, 2017, 8(3): 277-291.

[21] 杨欣, 袁自然, 叶寅, 等. 基于无人机高光谱遥感的冬小麦全氮含量反演[J]. 光谱学与光谱分析, 2022, 42(10): 3269-3274. [Yang Xin, Yuan Ziran, Ye Yin, et al. Winter wheat total nitrogen content estimation based on UAV hyperspectral remote sensing[J]. Spectroscopy and Spectral Analysis, 2022, 42(10): 3269-3274.]

[22] Miao J, Zhen J N, Wang J J, et al. Mapping seasonal leaf nutrients of mangrove with Sentinel-2 images and XGBoost method[J]. Remote Sensing, 2022, 14(15): 3679, doi: 10.3390/rs14153679.

[23] Komolafe A A, Olorunfemi I E, Oloruntoba C, et al. Spatial prediction of soil nutrients from soil, topography and environmental attributes in the northern part of Ekiti State, Nigeria[J]. Remote Sensing Applications: Society and Environment, 2021, 21: 100450, doi: 10.1016/j.rsase.2020.100450.

[24] 代子俊. 近30年湟水流域土壤养分时空变异及影响因素[D]. 西宁: 青海师范大学, 2018. [Dai Zijun. Spatio-temporal variation of soil nuterents in Huangshui River Basin and its affecting factors in the past 30 years[D]. Xining: Qinghai Normal University, 2018.]

[25] 方慧婷, 蒙继华, 程志强. 基于遥感与作物模型的土壤速效养分时空变异分析[J]. 中国农业科学, 2019, 52(3): 478-490. [Fang Huiting, Meng Jihua, Cheng Zhiqiang. Spatio-temporal variability of soil available nutrients based on remote sensing and crop model [J]. Scientia Agricultura Sinica, 2019, 52(3): 478-490.]

[26] 刘尊方, 雷浩川, 雷蕾. 湟水流域土壤有机质和速效磷空间布局分析[J]. 科学技术与工程, 2022, 22(34): 15095-15102. [Liu Zunfang, Lei Haochuan, Lei Lei. Analysis on spatial distribution of soil organic matter and available phosphorus in Huangshui River Basin[J]. Science Technology and Engineering, 2022, 22(34): 15095-15102.]

[27] Chen B L, Zheng H W, Luo G P, et al. Adaptive estimation of multi-regional soil salinization using extreme gradient boosting with Bayesian TPE optimization[J]. International Journal of Remote Sensing, 2022, 43(3): 778-811.

[28] 胡亚男, 高小红, 申振宇, 等. 基于野外实测Vis-NIR光谱的土壤肥力估算研究——以湟水流域为例[J]. 土壤通报, 2021, 52(3): 575-584. [Hu Ya'nan, Gao Xiaohong, Shen Zhenyu, et al. Estimating fertility index by using field-measured Vis-NIR spectroscopy in the Huangshui River Basin[J]. Chinese Journal of Soil Science, 2021, 52(3): 575-584.]

[29] Ji W S, Liu Y Q. Research on quantitative evaluation of remote sensing and statistics based on wireless sensors and farmland soil nutrient variability[J]. Computational Intelligence and Neuroscience, 2022, 2022: 3646264, doi: 10.1155/2022/3646264.

[30] 郭静, 龙慧灵, 何津, 等. 基于Google Earth Engine和机器学习的耕地土壤有机质含量预测[J]. 农业工程学报, 2022, 38(18): 130-137. [Guo Jing, Long Huiling, He Jin, et al. Predicting soil organic contents in cultivated land using Google Earth Engine and machine learning[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(18): 130-137.]

[31] 胡贵贵, 杨粉莉, 杨联安, 等. 基于主成分和机器学习的土壤有机质含量空间预测建模[J]. 干旱区地理, 2021, 44(4): 1114-1124. [Hu Guigui, Yang Fenli, Yang Lian'an, et al. Spatial prediction modeling of soil organic matter content based on principal components and machine learning[J]. Arid Land Geography, 2021, 44(4): 1114-1124.]

[32] 李莹莹, 赵正勇, 杨旗, 等. 基于GF-1遥感数据预测区域森林土壤有机质含量[J]. 土壤, 2022, 54(1): 191-197. [Li Yingying, Zhao Zhengyong, Yang Qi, et al. Prediction of soil organic matter content based on artificial neural network model and GF-1 remote sensing data[J]. Soils, 2022, 54(1): 191-197.]

[33] 孙铭岳. 基于多源遥感的黄河三角洲耕地土壤盐分和有机质含量反演[D]. 泰安: 山东农业大学, 2022. [Sun Mingyue. Inversion of soil salinity and organic matter on arable land in the Yellow River Delta based on multi-source remote sensing[D]. Tai'an: Shandong Agricultural University, 2022.]

[34] 杜军, 古军伟, 邱士可, 等. 基于支持向量回归的土壤全氮含量高光谱估测研究[J]. 河南科学, 2020, 38(10): 1585-1590. [Du Jun, Gu Junwei, Qiu Shike, et al. Prediction of total nitrogen by using hyperspectral data based on support vector regression[J]. Henan Science, 2020, 38(10): 1585-1590.]

Remote sensing inversion of soil nutrient on farmland in Huangshui River Basin based on XGBoost model

LIU Zunfang[1], LEI Haochuan[1], SHENG Haiyan[2]

(1. Department of Geological Engineering, Qinghai University, Xining 810016, Qinghai, China;
2. College of Agriculture and Animal Husbandry, Qinghai University, Xining 810016, Qinghai, China)

Abstract: The Huangshui River Basin is an important part of the Huangshui Valley. Additionally, collaborative environmental factors that predict the spatial distribution of soil nutrients are particularly important for managing soil nutrients. Moreover, less attention is paid to the effect of model parameters on the results obtained from soil nutrient inversion studies. In this study, the Huangshui River Basin in Qinghai Province (China) was selected as the study area, and 28 factors, including elevation, aspect, slope, plane curvature, section curvature, relief degree of land surface, topographic wetness index, soil pH, and spectral reflectance, were selected. In addition, these factors were combined with the Bayesian optimization algorithm (BOA) to construct artificial neural network (ANN), support vector machine (SVM), and extreme gradient boosting (XGBoost) machine learning models for predicting the spatial distribution of four soil nutrients in farmlands: soil organic matter (SOM), total nitrogen (TN), available phosphorus (AP), and available potassium (AK), respectively. Further, the prediction accuracy of these three models was evaluated based on the model coefficient of determination (R^2), root-mean-square error (RMSE), and relative percent deviation (RPD). The results revealed that: (1) All four soil nutrients exhibited a moderate degree of variability, with TN showing the highest variability of 69.481%. The XGBoost model based on the Bayesian optimized hyperparameter combination was better than other models in predicting the TN content (R^2, RMSE, and RPD were 0.893, 0.359, and 2.470, respectively). The R^2 values of the XGBoost model validation set for estimating the SOM, AK, and AP contents were 0.801, 0.509, and 0.442, respectively, and the corresponding RPD values were 2.152, 1.210, and 1.274, respectively. Moreover, this model exhibited a better prediction capability. (2) The comparison of the number of optimizations and errors of the three models revealed that the BOA-XGBoost model exhibited minimum number of parameter optimizations, higher efficiency, and better robustness. The ANN and SVM models demonstrated different prediction accuracies for different nutrients; additionally, the SVM model predicted the SOM content with high accuracy (RPD=1.580), while the ANN model predicted TN efficiently (RPD=2.460). Based on Landsat 8 remote sensing images, the XGBoost inversion model developed by combining 28 factors of the Huangshui River Basin was found to be more suitable for application in soil nutrient inversion research; furthermore, it can more accurately describe the spatial distribution pattern of the soil nutrient inversion in the Huangshui River Basin, better ensure precise agriculture fertilization, improve the fertilizer utilization rate and crop yield, and provide a reference for precise agriculture fertilization in the Huangshui River Basin.

Key words: soil nutrient; XGBoost; spatial distribution; environmental factor; Huangshui River Basin